caBIG Workflow

Brian Gilman, Panther Informatics

Patrick McConnell, Duke

William Sanchez, SAIC

Shannon Hastings, OSU

Agenda

1. Introduction to workflow (Brian Gilman)
   - HapMap, proteomics, microarray use cases
2. Review of technologies (Patrick McConnell)
   - BPEL, WSCI, Taverna, GenePattern, Pegasus, PAWS
3. Workflow architecture (William Sanchez)
   - Grid fabric, services, interfaces, workflow
4. Issues facing caBIG (Shannon Hastings)
   - Users, simplicity/flexibility/standards, provenance
5. Recommendations (Patrick McConnell)
   - White paper, reference implementations

Introduction to workflow

Brian Gilman, Panther Informatics

Use case 1: Haplotype Map Project

- GOAL: to define blocks and identify haplotypes throughout the genome in order to make scans for medically relevant variation feasible
- international collaboration in the spirit of the human genome project
- at least three populations to be examined

Potential impact of genetics on medical practice

- Understand fundamental basis of disease
- Risk-stratify patients for preclinical intervention
- Predict outcome and likely response to treatment

Human genetic variation

- To date, more than 4 million SNPs discovered through comparison of two “copies” of pieces of the genome
- 90% of this variation is common (>1%)
- 10 million such “common” SNPs are expected to exist in the genome

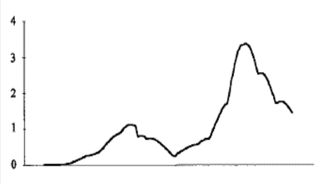
International SNP Map Working Group, Nature 2001

HYPOTHESIS:
Common Variation / Common Disease

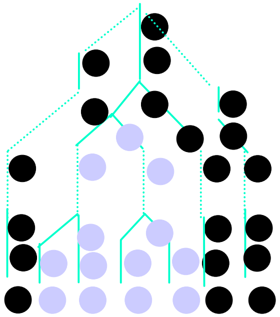
Since common polymorphisms (mostly SNPs) comprise the vast majority of variation among humans, it is very likely that they control the majority of common genetically influenced phenotypic variation (including risk to common disease)

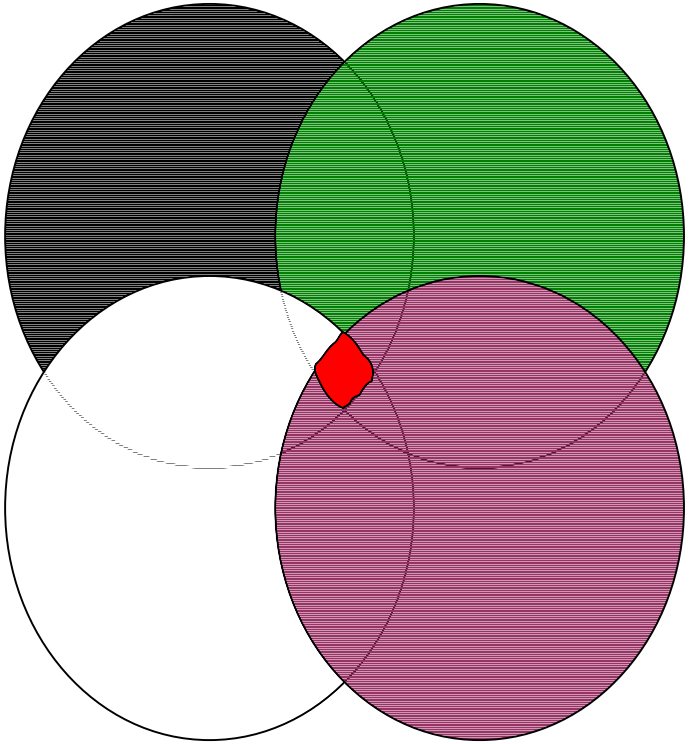
What will it take? Integrating information

Linkage analysis

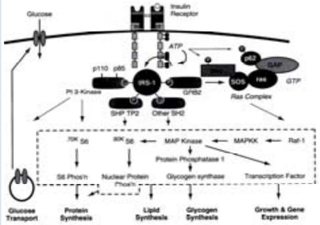


Haplotype mapping



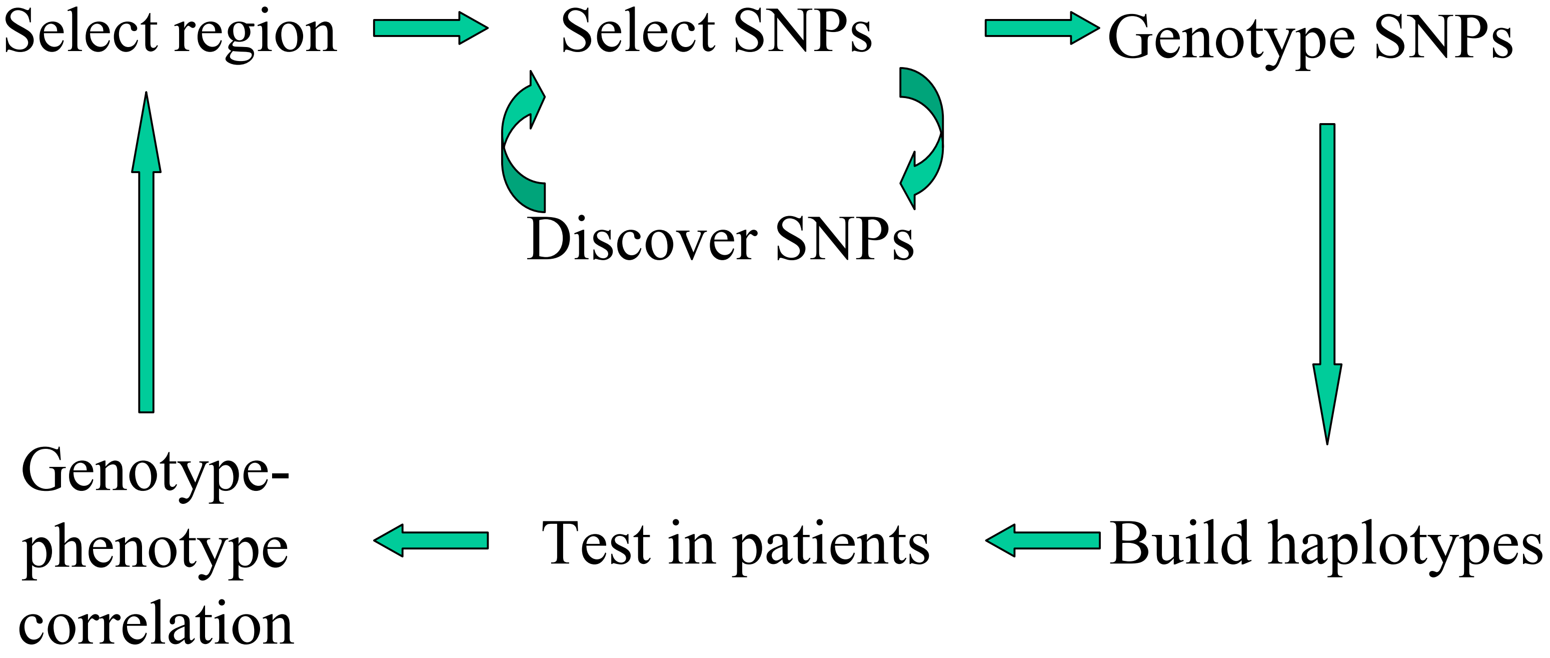


Expression monitoring



Biological pathways

Complex projects and datasets



Select region

Select SNPs

Genotype SNPs

Discover SNPs

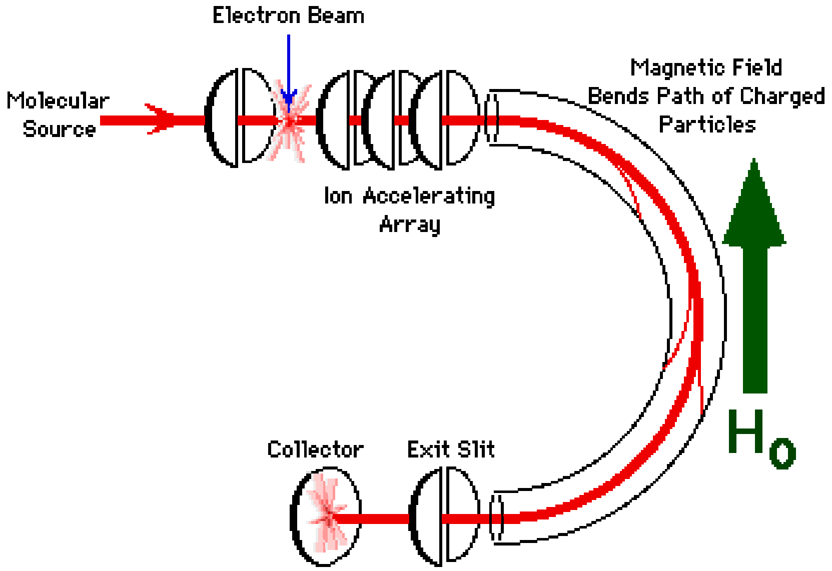
Build haplotypes

Test in patients

Genotype-phenotype correlation

Use case 2: Proteomics mass spectrometry

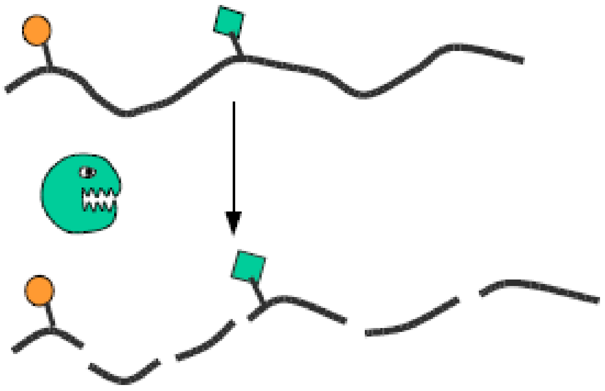
- Goal: identify or characterize proteins found in a biologic sample
- A substance is bombarded with an electron beam
- This breaks the substance into fragments of the original molecule
- These fragments are sorted by molecular weight to produce a spectrum



http://chipo.chem.uic.edu/web1/ocol/spec/MS1.htm

Proteomics identification studies

Tissue

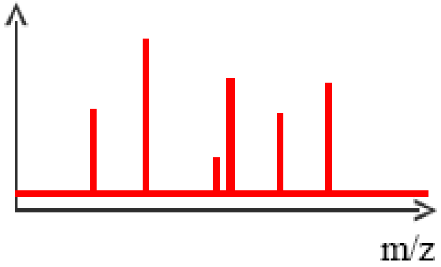


Protein digest database

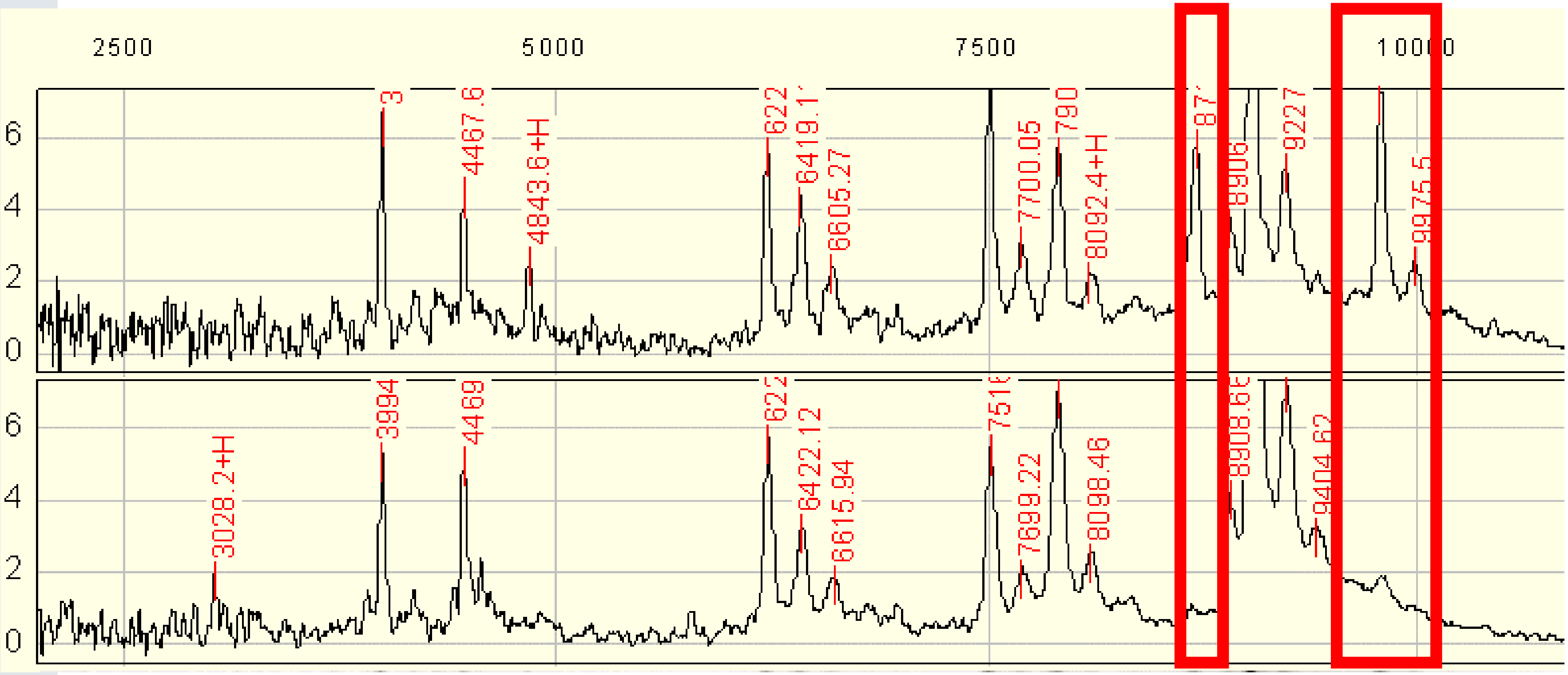
(MW of all tryptic peptides of all proteins)

MSPQTETK	922.05
ASVGFK	608.72
AGVK	374.46
EYK	439.49
LTYYTPEYETK	1408.55
DTDILAAFR	1022.15
VTPQPGVPPEE	3857.18
YK	310.37
GR	232.26
CYHIEPVPGEET	1996.23

Identify Protein

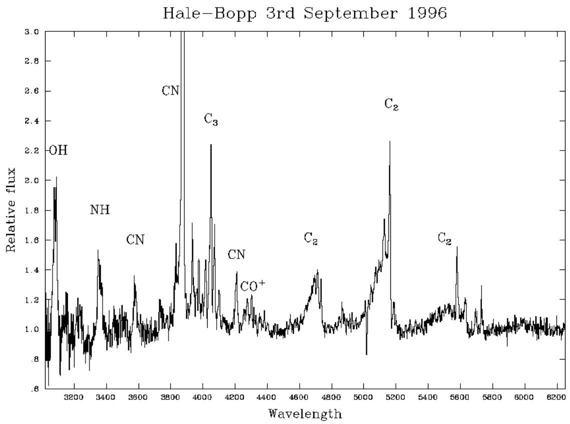


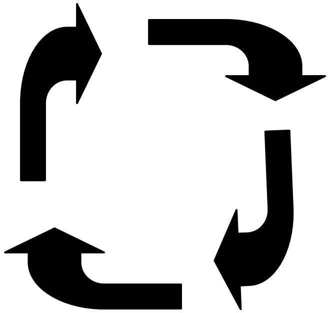
Proteomics profiling studies



Courtesy: Ciphergen

Proteomics analysis workflow





Smoothing
Interpolation

Peak finding

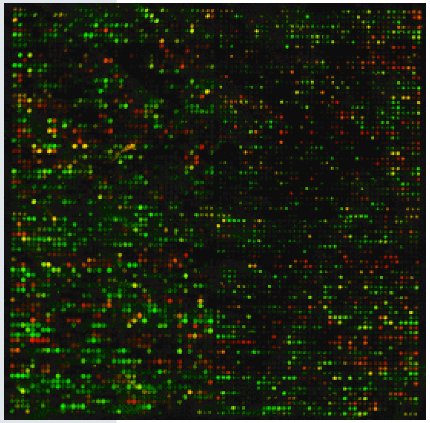
Data aggregation
Normalization

PCA
ICA

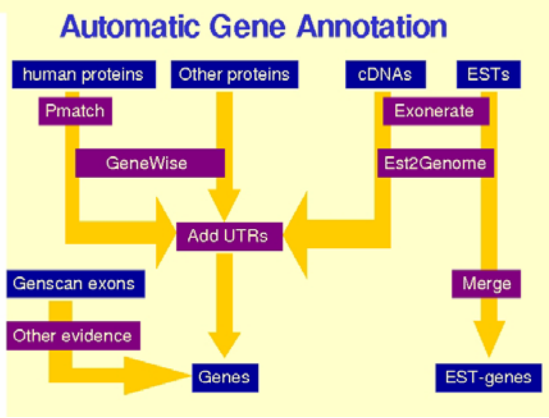
Supervised learning
Survival modeling

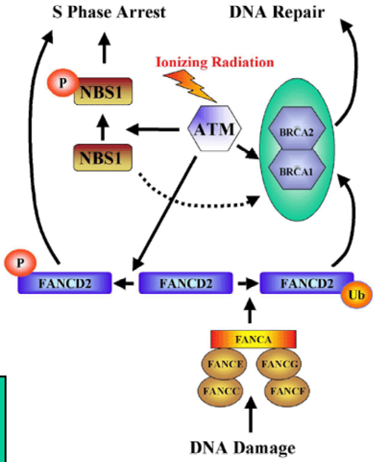
Source: Duke use cases

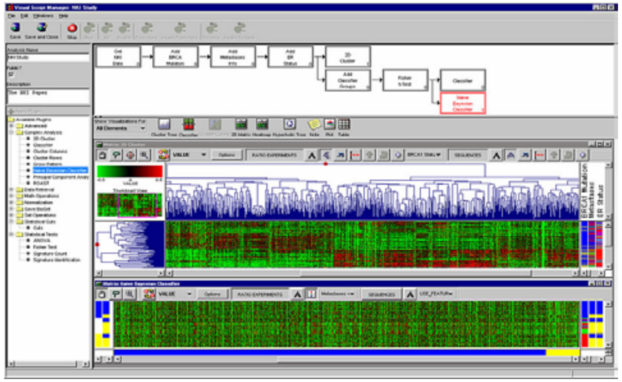
Use case 3: Microarray data/analysis workflow



Normalization
Baseline removal







Clustering

Source: Extension of Moffitt use cases

Review of technologies

Patrick McConnell, Duke

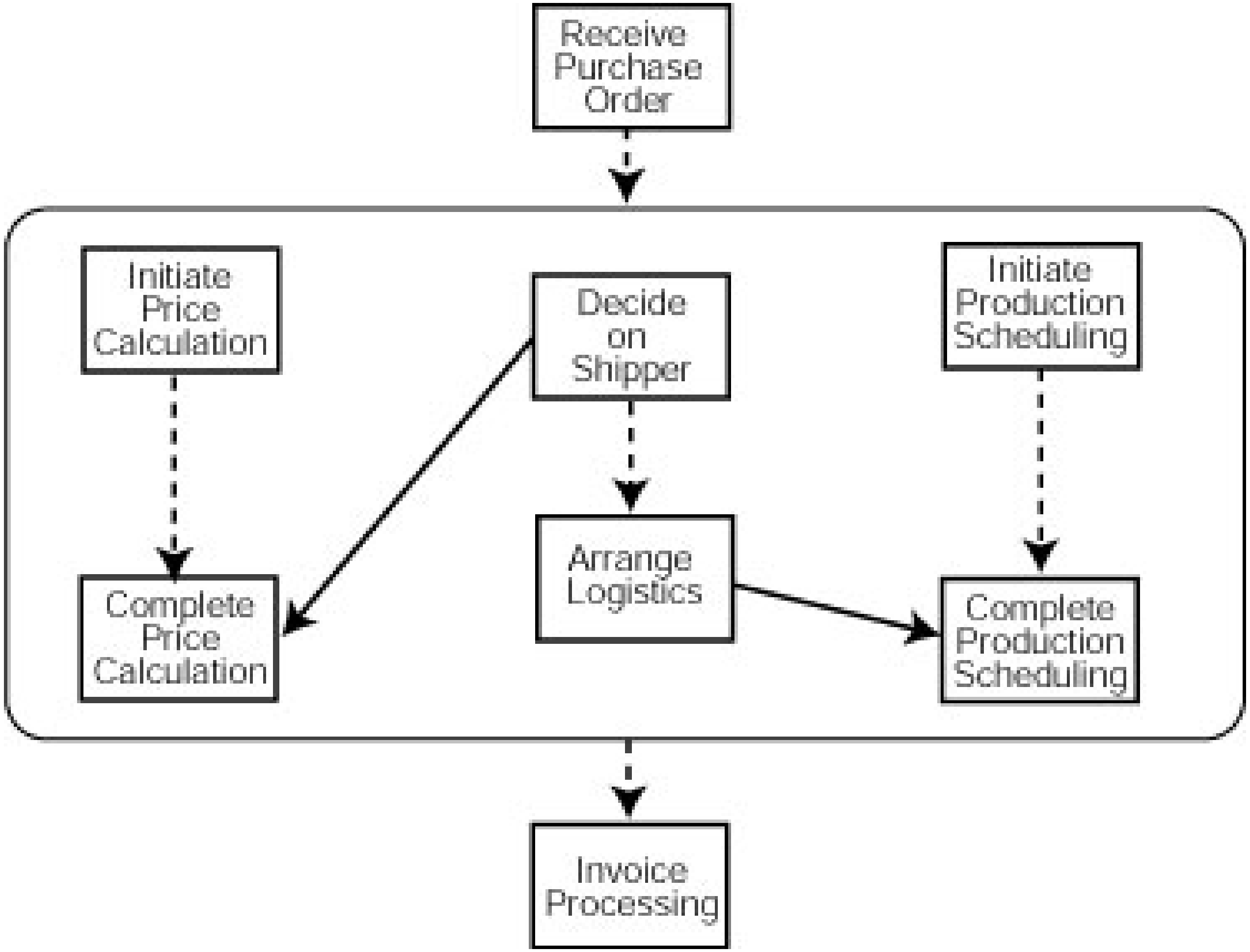
Workflow defined

- Orchestration
  - Describes interactions at message level
  - Business logic and execution order
  - Perspective of one of the business parties
- Choreography
  - More collaborative
  - Observable behavior between services
  - Tracks the sequence of messages
  - Perspective of the services

BPEL

- Business Process Execution Language
- Originally by Microsoft, IBM, Siebel Systems, BEA, and SAP
- Now being accepted by OASIS (read: standards)
- Implementation by IBM, Oracle
- Problems: complex, still not mature
- Benefits: standards-based

BPEL example

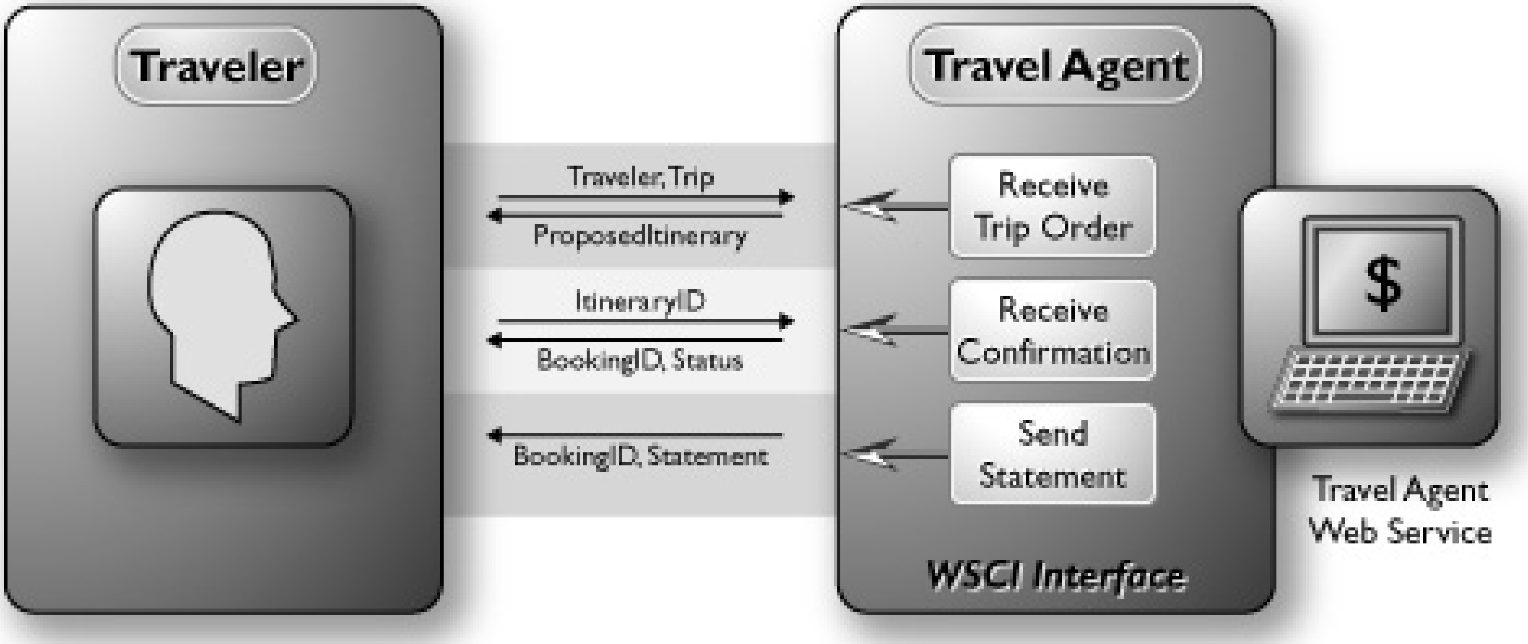

Receive Purchase Order
Initiate Price Calculation
Decide on Shipper
Initiate Production Scheduling
Arrange Logistics
Complete Price Calculation
Complete Production Scheduling
Invoice Processing

http://www-106.ibm.com/developerworks/library/ws-bpel/#initialexample

WSCI

- Web Services Choreography Interface
- Originally Sun, SAP, BEA, Intalio
- Now a W3C note
- “Not a workflow description language”
- “It can describe the observable behavior of a system that implements a workflow”
- Problems: no implementation yet
- Benefits: standards based

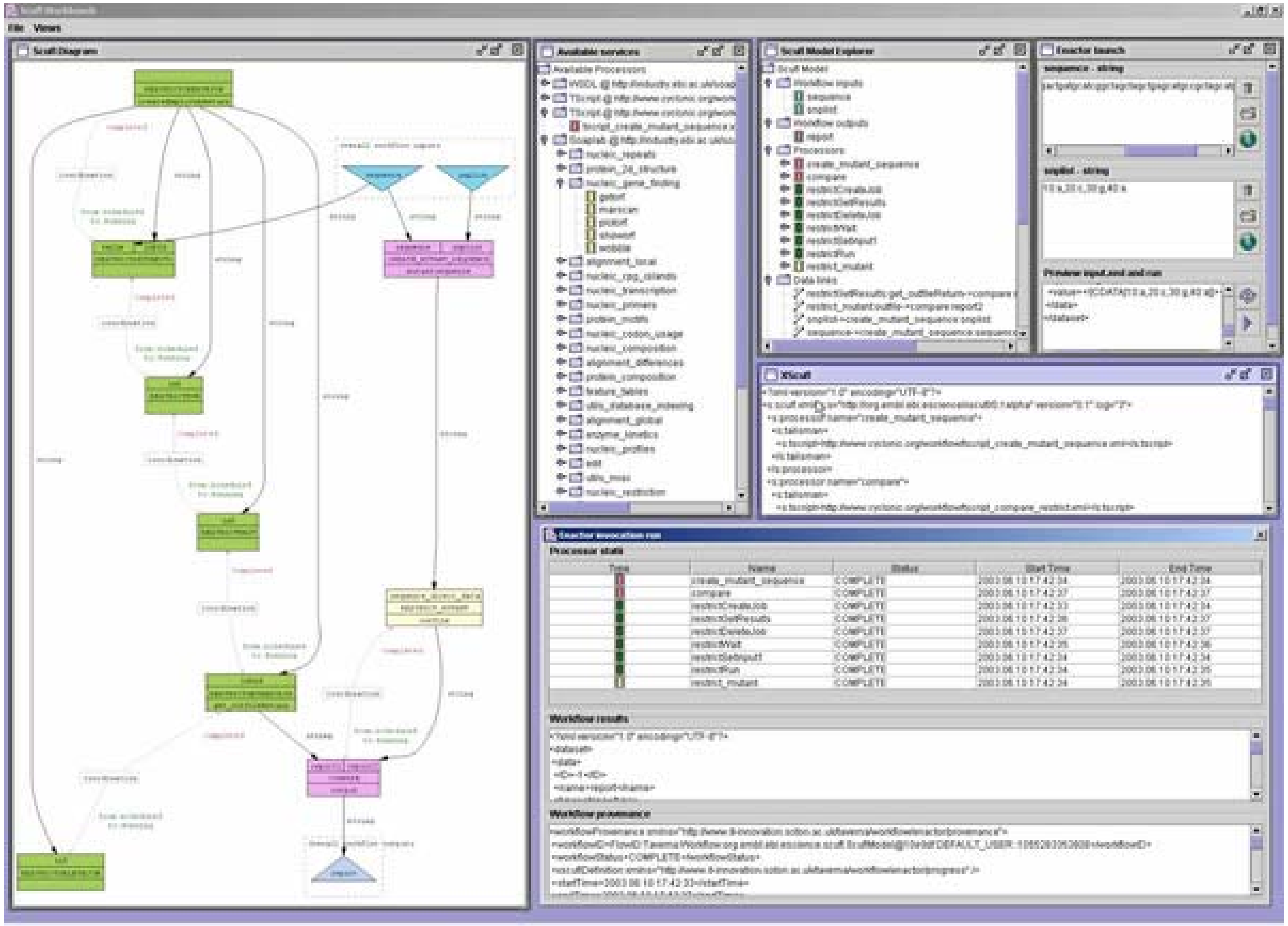
WSCI example



Taverna

- Open source project for bioinformatics workflows
- GUI-based with workflow engine behind
- Problems: overly simple
- Benefits: some community acceptance

Taverna screenshot



```
<?xml version="1.0" encoding="UTF-8"?>
<s:scufl xmlns:s="http://org.embl.ebi.escience/xscufl/0.1alpha" version="0.1" log="3">
  <s:processor name="create_mutant_sequence">
    <s:talisman>
      <s:tscript>http://www.cyclonic.org/workflows/tscript_create_mutant_sequence.xml</s:tscript>
    </s:talisman>
  </s:processor>
  <s:processor name="compare">
    <s:talisman>
      <s:tscript>http://www.cyclonic.org/workflows/tscript_compare_restrict.xml</s:tscript>
```

Enactor invocation run

Processor stats

Type	Name	Status	Start Time	End Time
🟥	create_mutant_sequence	COMPLETE	2003.06.10.17.42.34	2003.06.10.17.42.34
🟥	compare	COMPLETE	2003.06.10.17.42.37	2003.06.10.17.42.37
🟩	restrictCreateJob	COMPLETE	2003.06.10.17.42.33	2003.06.10.17.42.34
🟩	restrictGetResults	COMPLETE	2003.06.10.17.42.36	2003.06.10.17.42.37
🟩	restrictDeleteJob	COMPLETE	2003.06.10.17.42.37	2003.06.10.17.42.37
🟩	restrictWait	COMPLETE	2003.06.10.17.42.35	2003.06.10.17.42.36
🟩	restrictSetInput	COMPLETE	2003.06.10.17.42.34	2003.06.10.17.42.34
🟩	restrictRun	COMPLETE	2003.06.10.17.42.34	2003.06.10.17.42.35
🟨	restrict_mutant	COMPLETE	2003.06.10.17.42.34	2003.06.10.17.42.35

Workflow results

```
<?xml version="1.0" encoding="UTF-8"?>
<dataset>
<data>
 <ID>-1</ID>
 <name>report</name>
```

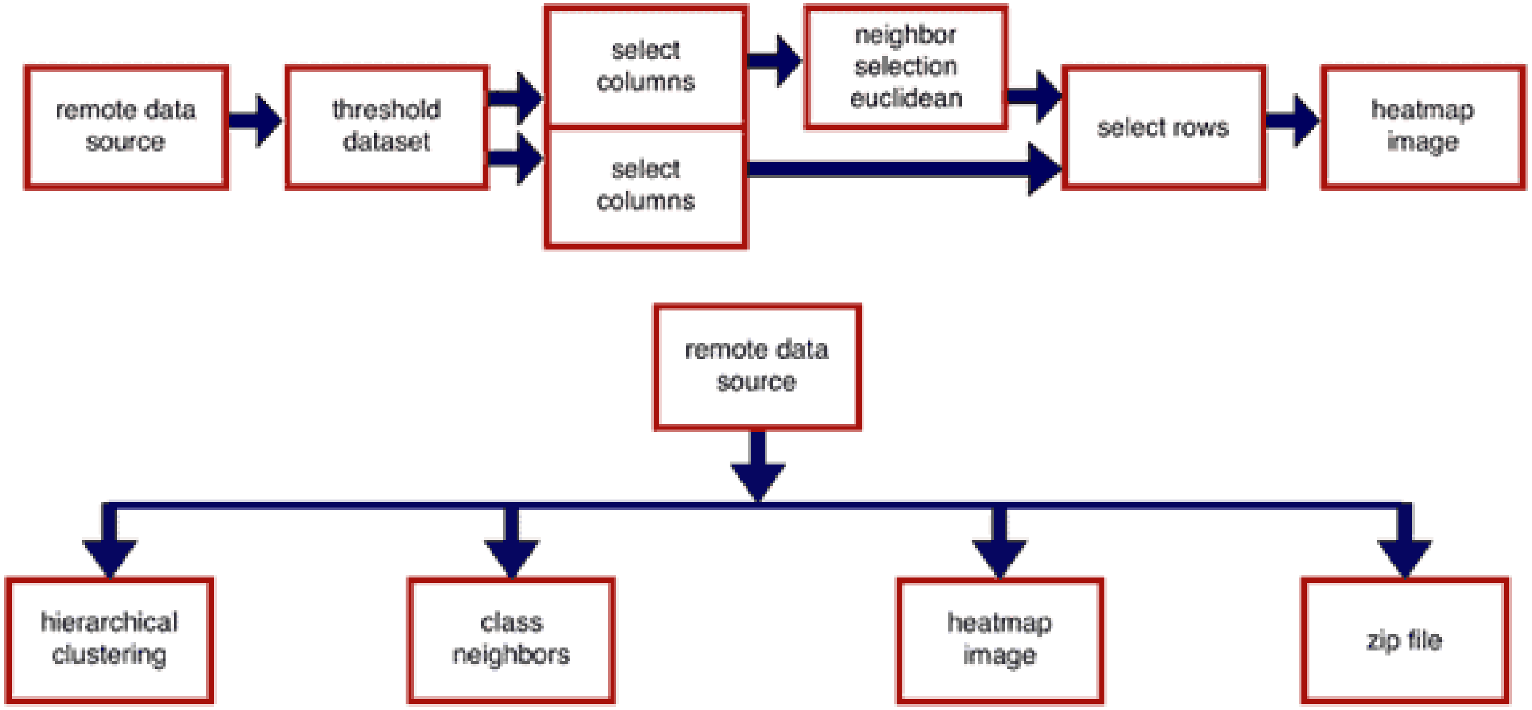
Workflow provenance

```
<workflowProvenance xmlns="http://www.it-innovation.soton.ac.uk/taverna/workflow/enactor/provenance">
<workflowID>FlowID:TavernaWorkflow.org.embl.ebi.escience.scufl.ScuflModel@[illegible]DEFAULT_USER:[illegible]</workflowID>
<workflowStatus>COMPLETE</workflowStatus>
<scuflDefinition xmlns="http://www.it-innovation.soton.ac.uk/taverna/workflow/enactor/progress" />
<startTime>2003.06.10.17.42.33</startTime>
```

GenePattern

- By MIT (ICR funded project)
- Platform for running pipelines of analyses
- Utilizes SOAP for client/server communication
- Also has
  - Data visualization environment
  - Analysis tool repository
  - Smooth integration with Java and R

GenePattern example



http://www.broad.mit.edu/cancer/software/genepattern/tutorial/

Pegasus

- Scientific workflow on the Grid
- Integrates with
  - Chimera (data derivation)
  - Condor (grid task engine)
  - MCS (Metadata Catalogue Service)
- Proprietary workflow language
- Maps abstract workflows to the Grid environment
- Has been applied to bioinformatics
  - Blast of 450 genomes (75GB)
  - 2D e. microscope -> 3D structures (200GB)

Pegasus example

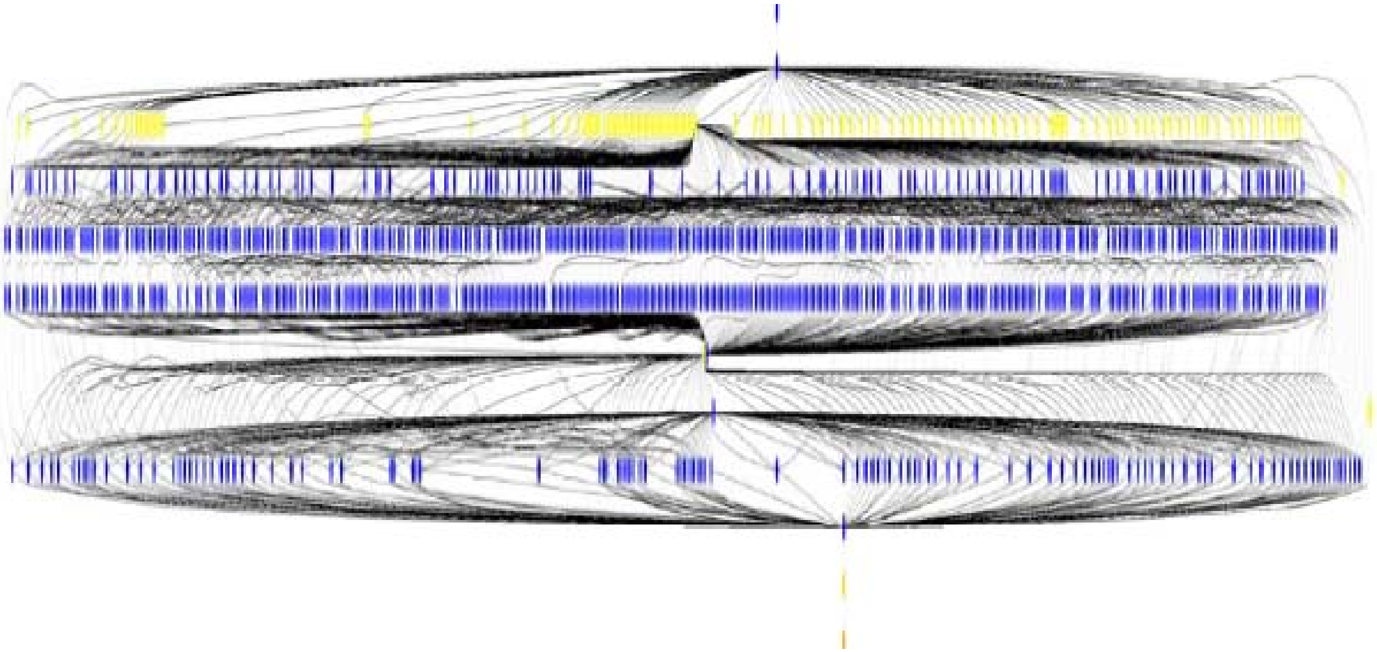


Figure 4: Montage workflow produced by Pegasus. The light colored-nodes represent data stage-in and the dark colored nodes, computation.

http://pegasus.isi.edu/

PAWS

- Panther Informatics Analysis Workflow System
- Bioinformatics workflow engine, language, dynamic generation of GUIs
- Pipeline-oriented (loosely based on OmniGene)
  - Transmission, Discovery, Analysis, Visualization, Relationship Management
- Open Source (primarily Java, but many other API's available)
- Language access web services, command-line tools, statistical routines, Java methods, etc.

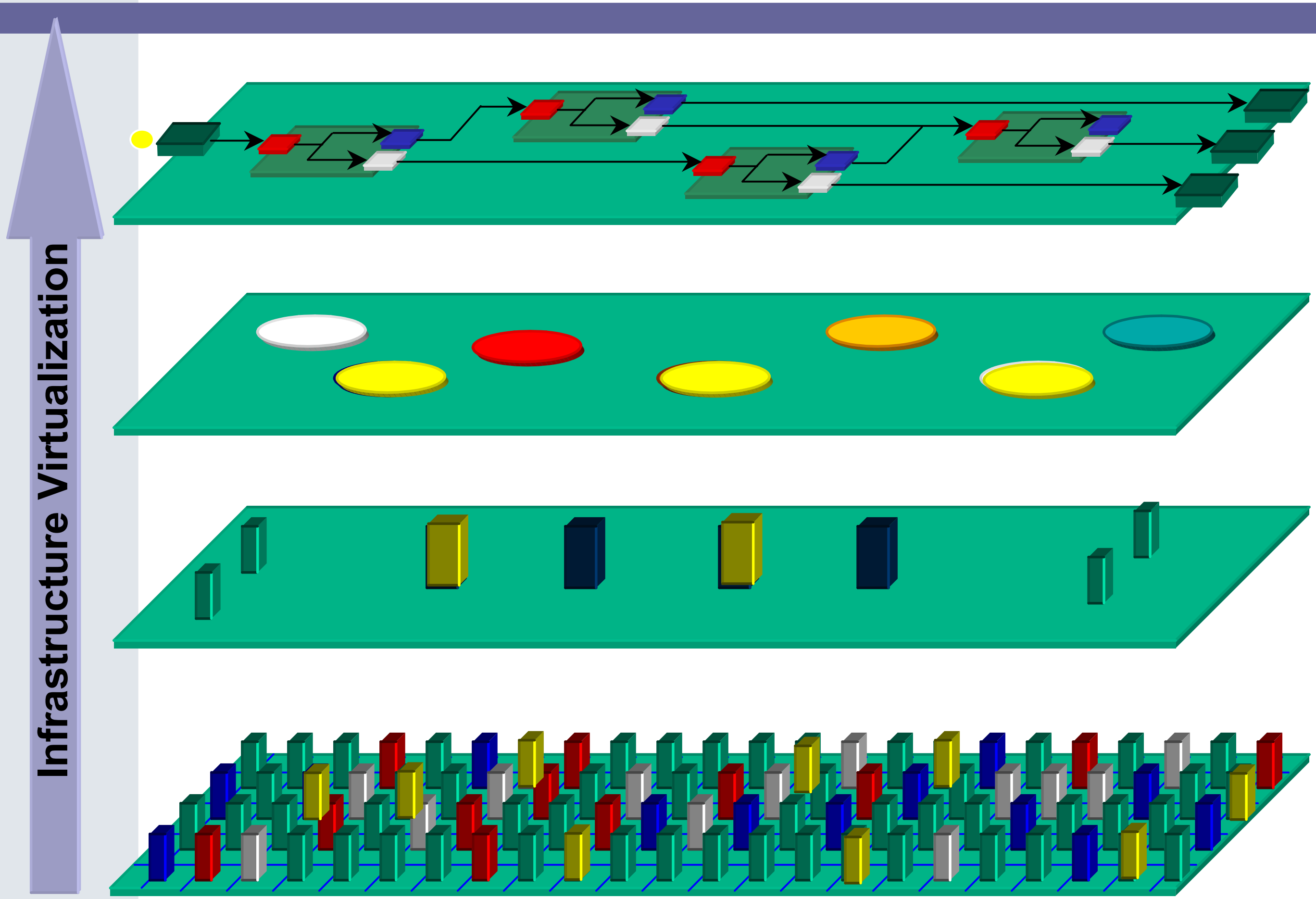
PAWS framework

Workflow architecture

William Sanchez, SAIC

Business Integration Services Grid

- Business service underlying implementation and deployment is dynamically managed by Grid Services infrastructure
- Service instances adapt to demand to provide QoS
- Service implementation details are abstracted from the flow engine
- Grid services infrastructure manages QoS: service completion times, guaranteed service completion and availability



Intra or Inter Enterprise Micro and Macro flows

- Service flow, choreography
- Service composition

Grid Services Interfaces

- Business functions or resources
- Instance registry, service data inspection
- QoS and cost descriptors

Grid Services Infrastructure

- Scheduling and monitoring
- Provisioning and deployment
- Usage metering
- Service and resource adapters

Grid Fabric

- Servers, clusters, farms, storage
- Bandwidth
- Inter or Intra Enterprise Data Centers
- Possibly geographically distributed

NATIONAL CANCER INSTITUTE

caBIG cancer Biomedical Informatics Grid

Issues facing caBIG

Shannon Hastings, OSU

Who are the users?

- Indirectly, just about everyone.
- Programmers, grid builders, will define workflows which compose services together to provide seamless grid applications.
  - Lower level users who will be savvy in the technology
- Domain experts and/or informaticists may use tools to help them define workflows for data curation, querying, analysis, and so on.
  - High level domain users who will want easy orchestration of services via simple domain specific GUI for example.

What is important about a workflow language?

- **Simple** and easy to use API and or language.
- **Flexible** so that we don't run into workflows which can't be expressed by the language and executed by the system.
- Excepted community **standard** so that external support can be leveraged.

Provenance Issues

- What information should be tracked?
  - Who made transformations?
  - What transformations where made?
  - When were transformations made?
  - Where were the transformations made?
- UPS analogy
  - A package is delivered to you by UPS.
    - Along the way the package is tracked at every location.
    - The handler?
    - The time? (Inbound and Outbound)
    - The location?
    - The action on the package? (scanned, packed, etc)
    - What the package looks like?
  - All critical information to the auditing and curation of the system.

Recommendations

Patrick McConnell, Duke

Recommendations

- White paper
  - Identify issues (data provenance, complexity, maturity, etc.)
  - Review relevant technologies
  - Provide informed recommendations
- Reference implementation(s)
  - Using the white paper, choose 1 or more technologies
  - Implement
    - Data workflow
    - Analytics workflow
    - Data + Analytics workflow